Journal of Contaminant Hydrology 120-121 (2011) 99–114

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

journal homepage: www.elsevier.com/locate/jconhyd

Optimal reconstruction of concentrations, gradients and reaction rates from particle distributions

D. Fernàndez-Garcia *, X. Sanchez-Vila

Department of Geotechnical Engineering and Geosciences, Technical University of Catalonia, Gran Capità s/n, 08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 14 December 2009
Received in revised form 20 April 2010
Accepted 5 May 2010
Available online 19 May 2010

Keywords:
Particle tracking
Reactive transport
Mixing
Heterogeneity
Subsurface hydrology

ABSTRACT

Random walk particle tracking methodologies to simulate solute transport of conservative species constitute an attractive alternative for their computational efficiency and absence of numerical dispersion. Yet, problems stemming from the reconstruction of concentrations from particle distributions have typically prevented its use in reactive transport problems. The numerical problem mainly arises from the need to first reconstruct the concentrations of species/components from a discrete number of particles, which is an error prone process, and then computing a spatial functional of the concentrations and/or its derivatives (either spatial or temporal). Errors are then propagated, so that common strategies to reconstruct this functional require an unfeasible amount of particles when dealing with nonlinear reactive transport problems. In this context, this article presents a methodology to directly reconstruct this functional based on kernel density estimators. The methodology mitigates the error propagation in the evaluation of the functional by avoiding the prior estimation of the actual concentrations of species. The multivariate kernel associated with the corresponding functional depends on the size of the support volume, which defines the area over which a given particle can influence the functional. The shape of the kernel functions and the size of the support volume determines the degree of smoothing, which is optimized to obtain the best unbiased predictor of the functional using an iterative plug-in support volume selector. We applied the methodology to directly reconstruct the reaction rates of a precipitation/dissolution problem involving the mixing of two different waters carrying two aqueous species in chemical equilibrium and moving through a randomly heterogeneous porous medium.

1. Introduction

Solute transport of conservative solutes has been traditionally studied using the Advection–Dispersion Equation (ADE) which is derived based on local arguments. In the last decades, though, this equation has been greatly questioned (e.g., Neuman and Tartakovsky, 2008, and references therein) and a number of non-local formulations have appeared as alternatives to describe effective transport of conservative species at intermediate distances. These formulations include Continuous Time Random Walks (CTRW) (Berkowitz et al., 2006), Fractional Advection–Dispersion Equations (fADE) (Benson et al., 2000), Multi-Rate Mass Transfer models (MRMT) (Haggerty and Gorelick, 1995) and memory functions (Carrera et al., 1998).

Irrespectively of the choice of the underlying Partial Differential Equation governing the problem, transport equations are solved by means of numerical methods. In this context, Random Walk Particle Tracking Methodologies (RWPT) constitute an attractive technique for their computational efficiency and absence of numerical dispersion. These methods simulate solute transport by tracking in time a large number of particles injected into the system, each one with a predefined associated mass. At any time step (with Δt being

* Corresponding author.
E-mail addresses: daniel.fernandez.g@upc.edu (D. Fernàndez-Garcia), xavier.sanchez-vila@upc.edu (X. Sanchez-Vila).

0169-7722/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jconhyd.2010.05.001

either constant or random), the particle moves due to the sum of two terms, one being deterministic (loosely associated to advective processes) and the other random (loosely associated to dispersive processes). For any given time, resident concentrations can be recovered by defining a grid and counting the mass per unit volume of the particles that are located within a given support volume. The concentration estimate of a given support is a random variable but eventually converges to the "true value" when the number of particles tends to infinity. In other words, the mean estimation error of this estimator of concentrations decreases with the number of particles. Despite these problems, particle tracking algorithms are a good alternative compared to other numerical methods such as finite elements or finite differences (e.g., Salamon et al., 2006a).

Particle tracking algorithms are a convenient alternative to study molecular diffusion problems. In such cases particles can be associated to molecules. This would mean that a mole of solute should be discretized into N_A (Avogrado's number, $N_A = 6.02\ 10^{23}$) molecules. In all practical applications found in the literature the number of particles used to discretized a given mass is more in the order of 10^4 to 10^7 (e.g., Riva et al., 2008). The traditional approach is based on considering that each particle actually represents a group of molecules. In this case particles have a representative size. This concept has been included into a family of methods named smoothed-particle hydrodynamics (SPH) (e.g., Tartakovsky and Meakin, 2005; Herrera et al., 2008), where a spatial distance (known as the "smoothing length") is defined, over which the properties of the particles are "smoothed" by a kernel function. This means that any physical quantity of any particle can be obtained by summing the relevant properties of all the particles which lie within the range of the kernel. Similarly to SPH methods, the approach actually taken by most of the existing particle tracking codes (e.g., Pollock, 1988; Salamon et al., 2006b) considers that particles have a zero support. All these methods share in common that they provide noisy estimates of the spatial or temporal distribution of concentrations. This is particularly disturbing when the objective is not actually estimating concentrations, but rather spatial or temporal derivatives, which is the case in most applications regarding stochastic hydrology or reactive transport.

The alternative we propose is based on considering that the particles being tracked are just a subsample of the population. This renders the problem of going from particles to concentrations to be just a reconstruction problem, where the spatial or temporal distribution of a given variable must be inferred from the observation of the spatial or temporal location of a relatively small subsample. The initial selection of the particles to be tracked (amongst all the possible ones) becomes an uncorrelated random sampling process, each particle/molecule being given the same probability. This way, the larger the concentration in a given volume, the larger the number of particles finally selected (assuming a sufficient number of particles is considered). Since dispersion processes are intrinsically random, a different realization of the solute transport problem can just be seen as coming from a different sample extraction from the initial population. The reconstruction problem considered only accounts for subsampling and therefore disregards the accuracy issues related to the flow problem solution (discretization errors and convergence).

The methods to carry out such reconstruction can be borrowed from many other problems in science. A widely used family of methods is based on kernel particle filters, where kernel density estimators (KDE) can be used to reconstruct the posterior PDF's of the variable of interest. Fields of application are signal processing, wireless communication, and robotics among others (see e.g., Chang and Ansari, 2005; Stoessel and Sagerer, 2006) and it is the approach taken in this paper. Particle filters can be considered a generalization of the traditional Kalman filtering methods. A compilation of different particle filtering methods with a comparison with Kalman filters and other types of grid-filtering methods can be found in a lucid discussion by Arulampalam et al. (2002).

In a KDE based approach, the multivariate kernel associated with the corresponding functional (e.g., the pdf of concentration estimator) depends on the size of the support volume, which defines the area over which a given particle can influence the functional. The shape of the kernel functions and the size of the support volume determines the degree of smoothing which is optimized to obtain the best unbiased predictor of the functional using an iterative plug-in support volume selector. Arguably, the main advantage of Kernel Density filters is its ability to directly reconstruct optimally not only concentrations, but also spatial and temporal derivatives by optimizing the kernel functions in order to properly reproduce all derivatives. The methodology, thus, mitigates the error propagation in the evaluation of the derivatives by avoiding the prior estimation of the actual concentrations of species. This results in a large reduction in the number of particles needed for proper map reconstruction, a reduction that is more significant for highly complex problems such as transport of reactive species in heterogeneous media.

Here, we present a simple and automatic KDE method to directly reconstruct the concentration gradients involved in transport quantities such as memory functions, mixing indexes and reaction rates, based on the estimation of marginal and conditional density functions of the concentration derivatives. The paper is structured as follows. First, Section 2 describes the target transport quantities to be estimated. Section 3 presents the theoretical framework and development of the KDE methodology. Finally, Section 4 shows an application of the method conducive to reconstruct the reaction rates of a precipitation/dissolution problem involving the mixing of two different waters in chemical equilibrium and moving through a homogeneous and a randomly heterogeneous porous medium. The homogeneous media is used to contrast the method against traditional approaches and analytical solutions. The heterogeneous media is used to illustrate the effect of mixing on reactive transport in a real field setting.

2. Background and motivation

There are a number of key hydrogeological problems where we are actually interested in directly obtaining a good and efficient estimator of the derivatives or gradients. In the last few years a great interest has arisen in the literature concerning memory functions. This concept is directly related to the slope of the breakthrough curve (see e.g.

Haggerty et al., 2000; Dentz and Berkowitz, 2003), s which can be written as

$$s = \frac{dlogc}{dlog\,t} = \frac{t}{c}\frac{dc}{dt}. \quad (1)$$

Spatial derivatives of concentration are key when the concept of dilution is introduced into the solute transport picture. Kitanidis (1994) defined the dilution index E, a measure of the degree of dilution in the system whose derivative is given in terms of the integral upon the full domain Ω of a combination of the spatial derivatives of concentrations through

$$\frac{dln\,E}{dt} = \int_\Omega \frac{1}{c}\nabla c^t \mathbf{D}\nabla c d\mathbf{x}. \quad (2)$$

The concept of dilution is directly related to that of concentration variance, σ_c^2, defined by Kapoor and Gelhar (1994), Kapoor and Kitanidis (1998). Remarkably, σ_c^2 follows an ADE type with a source term (or destruction term), $f_{\sigma_c^2}$, given by

$$f_{\sigma_c^2}(\mathbf{x},t) = 2\langle \nabla c'^t \mathbf{D}\nabla c'\rangle, \quad (3)$$

where $\langle\cdot\rangle$ denotes the expectation operator, and c' is the deviation of the concentration field with respect to the mean field, obtained by solving a macrodispersive ADE. Dilution is also the cause that the *pdf* (the probability density function) of concentrations displays a beta-type distribution (see, e.g., Fiorotto and Caroni, 2002; Bellin and Tonina, 2007). Actually, the *pdf* of concentrations in heterogeneous media can be defined as the ensemble mean of an auxiliary function Π (Sanchez-Vila et al., 2009) which itself follows an ADE type equation with a source term f_{pdf} given as

$$f_{pdf}(\mathbf{x},t) = -\frac{\partial^2 \Pi}{\partial c^2}\nabla c^t \mathbf{D}\nabla c. \quad (4)$$

Particle tracking approaches can also be applied to reactive transport problems. A particular case when reactive transport can be efficiently handled using particle tracking is the case when transport can be fully defined in terms of conservative quantities called components (see Fernàndez-Garcia et al., 2008, for an example). This is the case for example when all solutes are subject to the same velocity and dispersivity (at any given point) and reactions are instantaneous. In this case, and assuming that the vector of conservative quantities $\mathbf{u}$ follows an ADE, De Simoni et al. (2005) show that the vector of reaction rates per unit volume of fluid, $\mathbf{r}$, can be computed as

$$\mathbf{r}(\mathbf{x},t) = \mathbf{H}\nabla u^t \mathbf{D}\nabla u, \quad (5)$$

$\mathbf{H}$ being the Hessian matrix involving the second derivatives of concentration of solutes with respect to concentrations of components. This matrix is obtained from a chemical speciation process. Eq. (5) was further extended by Donado et al. (2009) to obtain reaction rates when the governing equation for the conservative quantities is a MRMT model. We note that there are other alternative approaches to model reactive transport with particle tracking avoiding the prior calculation of concentrations (e.g. Gillespie, 1977). Still, concentrations of reactants and products should be computed optimally afterwards.

From the non-exhaustive list presented it is clear that a number of problems in solute transport involve the computation of concentration gradients. Thus, it is crucial to find an optimal way to estimate these derivatives/gradients whenever a particle tracking approach is used to solve the underlying governing transport equation. The method we propose is capable of finding optimal estimations of the concentration gradients. It is true that some of the problems addressed involve the computation of products of derivatives. While it is true that the optimal of the product is not the product of optimals, we believe that our approach can be used even for these problems to provide suboptimal estimates of quantities such as Eqs. (1), (2) and (5).

3. Kernel density estimators

Particle tracking techniques produce discrete distributions of particle attributes (mass) that have to be converted to a continuous distribution of concentrations. The mathematical representation of the concentration field from particle distributions depends on the type of observation. Particle clouds observed at given times t_0 yield resident concentrations c_r (concentrations averaged over a support volume), whereas particles passing through control surfaces located at $\mathbf{x}_0$ (e.g., a pumping well) lead to flux concentrations c_f. By normalizing these concentrations we can define the following probability density functions,

$$p(\mathbf{x}) = \frac{\phi(\mathbf{x})c_r(\mathbf{x};t_0)}{\int_\Omega \phi(\mathbf{x})c_r(\mathbf{x};t_0)d\mathbf{x}}, \quad (6)$$

$$p(t) = \frac{Q(t)c_f(t;\mathbf{x}_0)}{\int_0^\infty Q(t)c_f(t;\mathbf{x}_0)dt}, \quad (7)$$

where $p(\mathbf{x})$ is the probability of finding a solute mass within the support volume $[\mathbf{x},\mathbf{x}+d\mathbf{x}]$ at a given time t, and $p(t)$ is the probability of finding a solute mass within the time interval $[t,t+dt]$ at a given control location. $\phi(\mathbf{x})$ is the porosity and $Q(t)$ is the flow rate at the outlet location. The particle mass is then related to the normalized concentrations by

$$p(\mathbf{x}) = \frac{m_p(\mathbf{X}_p)}{M_t}E\left\{\delta\left(\mathbf{x}-\mathbf{X}_p\right)\right\}, \quad (8)$$

$$p(t) = \frac{m_p(T_p)}{M_a}E\left\{\delta\left(t-T_p\right)\right\}, \quad (9)$$

where $E\{\cdot\}$ is the expectation operator over many injected random particles, $\mathbf{X}_p$ is the pth-particle location at time t, T_p is the first passage time of the pth-particle crossing the control surface, M_t is the total mass in the domain Ω at time t, and M_a is the total mass passing through the control surface,

$$M_t = \int_\Omega \phi(\mathbf{x})c_r(\mathbf{x};t_0)d\mathbf{x}, \quad (10)$$

$$M_a = \int_0^\infty Q(t)c_f(t;\mathbf{x}_0)dt. \quad (11)$$

A natural estimate of the probability density functions $p(t)$ or $p(\mathbf{x})$ is the relative frequency of mass, which basically consists in counting the particle mass falling into a given support. In the traditional approach to recover concentrations from particle distributions, this support is defined based on a given discretization of the domain in space and/or time so that $p(t)$ and $p(\mathbf{x})$ are only evaluated at the centroid of the discretization elements. This is mathematically written as

$$p(\mathbf{x}_j) \approx \hat{p}(\mathbf{x}_j) \equiv \frac{1}{M_t} \sum_p \frac{m_p I\{\mathbf{X}_p \in B_j\}}{\Delta V_j}, \tag{12}$$

$$p(t_j) \approx \hat{p}(t_j) \equiv \frac{1}{M_a} \sum_p \frac{m_p I\{T_p \in B_j\}}{\Delta t_j}, \tag{13}$$

where the overhat indicates the estimator, B_j is respectively the support volume or the time interval, $I\{\cdot\}$ is an indicator function defined as

$$I(\mathbf{x} \in B) = \begin{cases} 1 & \mathbf{x} \in B \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

and $\mathbf{x}_j$ and t_j denote the centroid of the jth discretization element in space and time with sizes ΔV_j and Δt_j, respectively. In the limit, for infinitesimal supports, $I/\Delta V_j$ and $I/\Delta t_j$ approach a δ-function. By definition, these estimators depend on the domain discretization, i.e., the choice of the support size ΔV_j and Δt_j, and the number and mass of particles falling into the support. In general, a small support combined with a finite number of particles leads to very noisy estimates, whereas an increase in the support tends to oversmooth (over or underestimate) the estimated concentration distribution. Thus, an optimal choice of the support size exists. In this context, kernel density estimators (KDE) provide a convenient mathematical framework to obtain this optimal support.

The KDE approach starts by generalizing the previous estimators as

$$p(\mathbf{x}) \approx \hat{p}(\mathbf{x}) \equiv \frac{1}{M_t} \sum_p m_p K_H(\mathbf{x} - \mathbf{X}_p), \tag{15}$$

$$p(t) \approx \hat{p}(t) \equiv \frac{1}{M_a} \sum_p m_p K_H(t - T_p), \tag{16}$$

where $K_H(\mathbf{x}-\mathbf{X}_p)$ and $K_H(t-T_p)$ are kernels or weighting functions dependent on the separation distance between the particle position/time and the point of estimation (e.g., Hardle, 1990). The traditional approach is recovered by using $K_H = I\{\cdot\}/\Delta V$, which is known as the box kernel function. The difference between the formulation given by Eqs. (12)–(13) and Eqs. (15)–(16) is that the averaging of the kernels can be done also at a point different from the discretization element centroid. Thus, albeit the traditional box model introduces discontinuities at the box edges, kernel estimators produces smooth functions of space and/or time. These kernel functions are weighting functions of the particle mass that defines its region of influence. They are usually defined to be symmetric density functions whose shape and size is parametrized based on a smoothing parameter, H. Thus,

$$\int K_H(\mathbf{s}) d\mathbf{s} = 1, \tag{17}$$

where the integral extends over the full domain. This parameter H defines how the particle attributes can influence the concentrations. For flux concentrations, K_H is a univariate distribution of particle arrival times and therefore H is a scalar parameter. For resident concentrations, $\mathbf{H}$ is a symmetric positive definite $d \times d$ matrix and K_H is a d-variate distribution of the particle position in space (d being the space dimension). The Kernel function can generally be expressed by means of elementary Kernel functions, K, defined as

$$K_H(\mathbf{s}) = \mathbf{H}^{-1/2} K(\mathbf{V}\mathbf{s}), \tag{18}$$

$$\mathbf{V} \cdot \mathbf{V}^t = \mathbf{H}. \tag{19}$$

The shape of the elementary Kernel functions, K and the choice of $\mathbf{H}$ determines the degree of smoothing of concentrations. A variety of Kernel functions can be used to generate smooth concentrations (e.g., Hardle, 1990). Well-known models are the Triangle model and the Gaussian model, respectively written as

$$K(\mathbf{s}) = \prod_{i=1}^{d} (1 - |s_i|) I\{|s_i| \le 1\}, \tag{20}$$

$$K(\mathbf{s}) = (2\pi)^{-d/2} exp\left(-\mathbf{s}^T \mathbf{s}\right). \tag{21}$$

Kernel density estimation (KDE) is a standard technique for exploring the histogram of unknown populations. The interest of KDE methods to particle tracking methodologies is manyfold: (1) Kernel estimates have desirable statistical properties as multimodal distributions (e.g., the presence of double peaks) as well as the identification of non-Fickian features of solute transport (e.g., pronounced tailing in breakthrough curves); (2) It directly provides not only good estimates of concentrations but also of their functionals (e.g., derivatives, gradients, dilution and corresponding indexes); and (3) It can be used to directly select the optimal degree of smoothing of the concentration and/or their derivatives based on data in an automatic way. The latter is briefly described in this section.

3.1. Optimal estimates of flux concentrations

Optimal estimates of $p(t)$ can be determined based on an appropriate selection of the smoothing parameter H, which is now a scalar parameter because the problem is one-dimensional. This is obtained by minimizing some error measure. A common choice is the Mean Integrated Square Error (MISE), defined as

$$MISE(H) = E\left\{\int (\hat{p}(t) - p(t))^2 dt\right\}. \tag{22}$$

Hereinafter, the limits of integration have been deliberately excluded in all integrals to denote that the domain where optimization is performed is a modeler's choice. As an example, one can be interested in a concentration value that tends asymptotically to a baseline value. Integrating up to infinity

leads to an unbounded integral. Thus, a cut-off must be imposed. Also, it would be possible to obtain different optimal estimates in prespecified subdomains. In the simulations presented in this paper we have used the full domain.

Since it is known that the choice of K has little effect on the behavior of $\hat{p}$, in Eq. (22) we avoided to write explicitly the dependence of the shape of K in *MISE*. The reason for this is that kernel functions can be rescaled such that the difference between two KDE estimates obtained using two different kernels is almost negligible (Marron and Nolan, 1989). *MISE* can be written as the sum of two terms, the integral of variance and the integrated squared bias,

$$MISE(H) = \int Var[\hat{p}(t)]dt + \int Bias^2[\hat{p}(t)]dt. \quad (23)$$

Assuming that all particles carry the same mass m_p, so that the total mass is $M = N_p m_p$, using a change of variable along with the Taylor expansion of $p(t)$ around t, and taking the limit of *MISE* as $N_p H \to \infty$ and $H \to 0$, the following asymptotic expressions can be found (e.g., Silverman, 1986; Hardle, 1990),

$$MISE(H) = AMISE(H) + O(H^4 + (N_p H)^{-1}), \quad (24)$$

$$AMISE(H) = \frac{H^4}{4} R(p''(t))(\mu_2(K))^2 + \frac{1}{HN_p} R(K), \quad (25)$$

where the first and second terms of *AMISE* are respectively the integrated squared bias and the asymptotic integral of variance of the estimator. $R(g)$ is the L^2-norm of a given function $g(t)$, defined as

$$R(g) = \int g(t)^2 dt, \quad (26)$$

and $\mu_n(g)$ is the n-th moment about the origin of $g(t)$, defined as

$$\mu_n(g) = \int t^n g(t) dt. \quad (27)$$

Hence, the minimization of *MISE* with respect to H calls for a compromise between oversmoothing (taking a large H to reduce the variance) and undersmoothing (taking a small H to reduce the bias). The optimal choice of the bandwidth, H^{opt}, is then obtained by disregarding higher order terms and minimizing the asymptotic expression of $MISE(H)$. Setting $dAMISE/dH = 0$ yields (Park and Marron, 1990)

$$H^{opt} = \left(\frac{R(K)}{R(p'')(\mu_2(K))^2 N_p} \right)^{1/5}. \quad (28)$$

Thus, it turns out that the optimal bandwidth H^{opt} is inversely proportional to the number of particles used to the power of 0.2 and depends on the unknown function $R(p'')$. In the limit, when $N_p \to \infty$ the optimal bandwidth consistently approaches to zero. $R(p'')$ needs to be further estimated. Several methods can be used for this matter: (1) rule-of-thumb methods (i.e., a guess of the reference distribution p); (2) cross-validation methods; and (3) plug-in methods. A review of these methods can be found in several papers (e.g., (Jones et al., 1996; Park and Marron, 1990)) and books (e.g., (Hardle, 1990)).

3.2. Optimal estimates of resident concentrations

Likewise for flux concentrations, optimal estimates of resident concentrations at time t, $p(\mathbf{x})$, can also be determined based on an appropriate selection of the support volume $\mathbf{H}$, defined based upon an integrated square error measure as

$$MISE(\mathbf{H}) = E\left\{ \int (\hat{p}(\mathbf{x}) - p(\mathbf{x}))^2 d\mathbf{x} \right\}. \quad (29)$$

In a one-dimensional problem, the estimation of resident concentrations is exactly the same as for flux concentrations exchanging t by x. Nonetheless, in practice, the spatial distribution of resident concentrations is evaluated in two or three dimensions. In this case, albeit the choice of the smoothing parameter is rather simple for univariate problems, it becomes difficult for multivariate distributions, where it is required to estimate several free parameters. Wand and Jones (1993) found that the choice of $\mathbf{H}$ in two-dimensional problems should, in general, account for the curvature and orientation of the true distribution. Moreover, its principal directions could not be chosen effectively using the sample covariance matrix.

Several approaches can be considered to estimate multivariate probability density functions (Wand and Jones, 1994; Sain, 2002; Duong and Hazelton, 2003; Duong et al., 2008). A simple and effective alternative to avoid the difficulties inherited in higher dimensions is through the use of marginal and conditional densities (Simonoff, 1995). Let us consider the estimation of resident concentrations in two dimensions. By the definition of the conditional density function,

$$p(x,y) = p(y|x)p(x) = p(x|y)p(y), \quad (30)$$

where $p(x|y)$ and $p(y|x)$ are the conditional density functions, and $p(x)$ and $p(y)$ are the marginal density functions. From Eq. (30), the bivariate density function can be estimated as (Simonoff, 1995)

$$p(x,y) = [p(y|x)p(x)p(x|y)p(y)]^{1/2}. \quad (31)$$

The marginal and conditional probability distributions are estimated from the following expressions

$$p(x) \approx \hat{p}(x) \equiv \frac{1}{M_t} \sum_p m_p K_H(x - X_p), \quad (32)$$

$$p(y) \approx \hat{p}(y) \equiv \frac{1}{M_t} \sum_p m_p K_H(y - Y_p), \quad (33)$$

$$p(x|y) \approx \hat{p}(x|y) \equiv \frac{1}{M_t(y)} \sum_p m_p K_H(x - X_p), \quad (34)$$

$$p(y|x) \approx \hat{p}(y|x) \equiv \frac{1}{M_t(x)} \sum_p m_p K_H(y - Y_p), \quad (35)$$

$$M_t(x) = \int \phi c(x,y) dy, \quad (36)$$

$$M_t(y) = \int \phi c(x,y) dx, \quad (37)$$

$$M_t = \int \int \phi c(x,y) dx dy, \quad (38)$$

where K_H is the kernel for univariate distributions ($d=1$). From this, we see that the problem of estimating multivariate distributions can be easily reduced to the simple estimation of the corresponding marginal and conditional univariate distributions. We refer to Simonoff (1995) for the details on the numerical implementation of Eq. (31).

3.3. Optimal estimates of the derivatives of flux concentrations

By knowing $p(t)$ and its derivative with respect to time we can determine the time derivative of the flux concentrations, i.e., the slope of a breakthrough curve, through

$$\frac{dc_f(t;\mathbf{x}_0)}{dt} = \frac{M_a}{Q(t)}\left(\frac{dp(t)}{dt} - p(t)\frac{dlnQ(t)}{dt}\right), \tag{39}$$

which is obtained by taking the derivative of Eq. (7). In this context, KDE methods are capable to provide a direct, data-based calculation of the optimal of the derivatives of $p(t)$ without having to use the estimates of concentrations $\hat{p}(t)$. The kernel density estimate $\hat{p}^{(k)}(t)$ of the kth derivative of a univariate probability density function $p(t)$ is (e.g., Engel et al., 1994)

$$\frac{d^k p(t)}{dt^k} \approx \hat{p}^{(k)}(t) \equiv \frac{1}{M_a}\sum_p m_p K_H^{(k)}\left(t - T_p\right), \tag{40}$$

where H denotes the support volume, and $K_H^{(k)}$ is defined based on its corresponding elementary kernel function

$$K_H^{(k)}(s) = \frac{1}{H^{1+k}} K_k\left(H^{-1}s\right), \tag{41}$$

$$K_k(s) = \frac{d^k K}{ds^k}\left(H^{-1}s\right), \tag{42}$$

that satisfies

$$\int K_k(s) s^j ds = \begin{cases} 0 & \text{for } j = 0,\ldots,k-1, k+1,\ldots,m-1 \\ (-1)^k k! & \text{for } j = k \\ \mu_m(K_k) \neq 0 & \text{for } j = m. \end{cases} \tag{43}$$

Eq. (42) establishes the relationship between the elementary kernel function of $\hat{p}^{(k)}(t)$ and that of $\hat{p}(t)$, and requires K to be kth differentiable. Using again the mean integrated squared error criterion to evaluate the expected error of the estimator, and assuming that all particles carry the same mass, the following optimal support can be derived (Engel et al., 1994),

$$H^{\text{opt}} = \left(\frac{(2k+1)R(K_k)m!^2}{2(m-k)(\mu_m(K_k))^2 R(p^{(m)})N_p}\right)^{1/(2m+1)}. \tag{44}$$

For completeness, the derivation of Eq. (44) is also given in Appendix A. The important point here is to note that the optimal support associated with the derivative of the density function $\hat{p}^{(k)}$ is not the same as for the density function $\hat{p}$. The fluctuations associated with the derivative of a density function are larger than those of the density function and, therefore, its corresponding optimal support should be increased to provide the same degree of smoothing.

3.4. Optimal estimates of the gradients of resident concentrations

Similar to the calculation of flux concentrations, by using Eq. (6), the gradients of the resident concentrations can be written as a function of $p(\mathbf{x})$,

$$\nabla c_r(\mathbf{x};t) = \frac{M_t}{\phi(\mathbf{x})}(\nabla p(\mathbf{x}) - p(\mathbf{x})\nabla ln\phi(\mathbf{x})). \tag{45}$$

Here, $\nabla p(\mathbf{x})$ is a function of space and therefore its estimation is difficult when the problem is not one-dimensional. From Eq. (30), and after some algebra,

$$\partial_x p(x,y) = sgn\left(\frac{dp(x|y)}{dx}\right)\left[p(y|x)\frac{dp(x)}{dx}\frac{dp(x|y)}{dx}p(y)\right]^{1/2} \tag{46}$$

$$\partial_y p(x,y) = sgn\left(\frac{dp(y|x)}{dy}\right)\left[p(x|y)\frac{dp(y)}{dy}\frac{dp(y|x)}{dy}p(x)\right]^{1/2} \tag{47}$$

where $sgn(z)$ is the signum function, whenever $z\neq 0$ then $sgn(z) = z/|z|$. Thus, the problem of estimating the gradients of the resident concentrations is reduced to the evaluation of the derivatives of univariate density functions, which is a well known problem. The calculation of the latter has been already described in Section 3.3 and in Appendix A.

4. Computational investigations

Particle tracking simulations of a fairly complex multispecies reactive system were conducted in a synthetic aquifer to illustrate the process of reconstructing the reaction rates of chemical species from particle mass distributions. The discussion is organized as follows: Firstly, the reactive transport problem is described in Section 4.1. Then, we examine the performance of the proposed KDE method by contrasting particle tracking solutions of the reaction rates taking place in a homogeneous medium against its corresponding analytical solution (Section 4.2). Finally, in Section 4.3, we explore the effects of heterogeneity and the choice of the transport model on the behavior of the reactive system.

4.1. Reactive transport problem

We consider a dissolution/precipitation problem involving the mixing of two different waters. Each water carries in solution two aqueous species, B_1 and B_2, in instantaneous local equilibrium with a solid mineral M_3. The corresponding reaction is

$$\nu_1 B_1 + \nu_2 B_2 \rightleftarrows M_3. \tag{48}$$

Without loss of generality we consider $\nu_1=\nu_2=1$. The law of mass action implies that the activities of both aqueous species, $\{B_i\}$ $(i=1,2)$, must satisfy the following condition

$$K_{eq} = \{B_1\}\{B_2\}, \tag{49}$$

where K_{eq} is the equilibrium constant. Assuming that the solution is diluted, then we can assume unit activity coefficients and reformulate Eq. (48) in terms of concentrations

$$K_{eq} = c_1 c_2. \tag{50}$$

In this particular transport problem, the mixing of any two waters in equilibrium with the mineral leads to oversaturation of the resulting mixture. Precipitation then takes place instantaneously in order to re-equilibrate the system.

The migration of the two aqueous species is described by means of a mass transfer model. Mass transfer models are capable to simulate a large variety of processes, including composite diffusive processes (Haggerty and Gorelick, 1995), slow advection (Willmann et al., 2008), and subgrid model heterogeneity (Fernàndez-Garcia et al., 2009). Here, we chose the widely used single-rate mass transfer model. This model considers an overlapped continuum media formed by a mobile domain, where advection–dispersion takes place, and one immobile domain, where mass can be transferred to and temporarily be trapped. Thus, the partial differential equations governing the chemical species are written as

$$\phi_m \frac{\partial c_i}{\partial t} = -\nabla\cdot(\mathrm{q}c_i - \phi_m \mathrm{D}_d \nabla c_i) - \phi_m \Gamma_i - \phi_m r_i - f_i, i = 1,2 \tag{51}$$

subject to the corresponding boundary and initial conditions, where ϕ_m [dimensionless] is the mobile porosity, c_i [ML^{-3}] is the volume-averaged concentration of the ith reactive specie, D_d [L^2T^{-1}] is the hydrodynamic dispersion coefficient tensor, and q [LT^{-1}] is the fluid velocity. The functions r_i, f_i, and Γ_i [$ML^{-3}T^{-1}$] are source/sink terms accounting for the mass removal due to chemical reactions, extractions, and mass transfer processes. The source/sink term Γ_i is

$$\Gamma_i(\mathrm{x},t) = \beta \frac{\partial c_{im,i}}{\partial t}, i = 1,2, \tag{52}$$

which describes the mass exchange between the mobile and the immobile domain of the ith specie (per unit of aquifer volume), β [dimensionless] is the field capacity of the immobile domain, and $c_{im,i}$ [ML^{-3}] is the immobile resident concentration of the ith specie. The mass transfer equation needed to close Eq. (51) together with Eq. (52) is written as

$$\frac{\partial c_{im,i}}{\partial t} = \alpha\left[c_i - c_{im,i}\right] - r_{im,i}, i = 1,2, \tag{53}$$

where the reaction rate $r_{im,i}$ is the source/sink term [$ML^{-3}T^{-1}$] that accounts for the solute removed from the immobile domain by precipitation (i.e., precipitated mass per unit of fluid volume and time), and α [T^{-1}] is the first-order mass transfer coefficient.

We assume that both species sample the same advective process (a counterexample would be colloidal material or sorptive solutes) and both have the same dispersion coefficient. Then, following (De Simoni et al., 2005) and Willmann et al. (2009), this system can be fully defined by means of conservative concentration components, defined as

$$u = c_1 - c_2, u_{im} = c_{im,1} - c_{im,2}. \tag{54}$$

Subtracting the two mass balance equations governing c_1 and c_2, the conservative components u and u_{im} follow conservative equations, i.e., without the presence of r_i and f_i terms in Eq. (53). Once u and u_{im} are obtained from solving the conservative transport problem subject to initial and boundary conditions, the species concentrations in the mobile and immobile domains can be explicitly computed by speciation as

$$c_i = (-1)^{i-1}\frac{u}{2} + \frac{1}{2}\left(u^2 + 4K_{eq}\right)^{1/2}, i = 1,2, \tag{55}$$

$$c_{im,i} = (-1)^{i-1}\frac{u_{im}}{2} + \frac{1}{2}\left((u_{im})^2 + 4K_{eq}\right)^{1/2}, i = 1,2. \tag{56}$$

Applying the chain rule to Eq. (53) and assuming that K_{eq} is uniform in space and time, Willmann et al. (2009) found the following expression of the reaction rates taking place at the mobile and immobile domains,

$$r_m(u) = f_{chm}(u) f_{mix}(u) \tag{57}$$

$$r_{im}(u_{im}) = f_{chm}(u_{im}) f_{mix}(u_{im}) \tag{58}$$

where

$$f_{chm}(u) = \frac{\partial^2 c_2}{\partial u^2} = \frac{2K_{eq}}{\left(u^2 + 4K_{eq}\right)^{3/2}} \tag{59}$$

$$f_{mix}(u) = \nabla^t \mathrm{D}\nabla u \tag{60}$$

In Eq. (58), $f_{mix}(\cdot)$ term is a measure of mixing, while $f_{chm}(\cdot)$ term is directly associated with the chemistry of the system and has an explicit expression in terms of the conservative concentration components. Total reaction rates can then be computed as

$$r = \frac{\phi_m}{\phi_{tot}} r_m + \frac{\phi_{im}}{\phi_{tot}} r_{im}, \tag{61}$$

where r is the total reaction rate (i.e., total precipitated mass per unit of fluid volume and time), ϕ_{tot} is the total porosity, $\phi_{tot}=\phi_{im}+\phi_m$, and r_m and r_{im} are respectively the reaction rates in the mobile and immobile domain.

4.2. Homogeneous medium

In order to test the methodology, we compare the f_{mix} estimates obtained with the proposed KDE method with an analytical solution. For this purpose, we consider a two-dimensional homogeneous aquifer under steady-state uniform flow conditions. Solute transport is described by the traditional reactive advection–dispersion equation, without external forces or the presence of an immobile domain, i.e., $r_i \neq 0$, $f_i=0$ and $\Gamma_i=0$ in Eq. (51). Initially, the two aqueous species $\{c_1, c_2\}$ are in local equilibrium with the solid mineral such that $c_{1,0}=c_{2,0}$ (i.e., $u_0=0$). The initial equilibrium is

then affected by a point-like instantaneous injection of water, that is still in chemical equilibrium with the mineral but with a different chemical composition. In this situation, knowing that $u = c_1 - c_2$ is the conservative specie (De Simoni et al., 2005), the initial condition can be written as

$$u\left(x, y, t = 0^{+}\right) = u_0 + \Delta u_0 \delta(x - x_0)\delta(y - y_0), \tag{62}$$

where $\delta(\cdot)$ is the Dirac delta function, (x_0,y_0) is the point of injection, and Δu_0 is the initial pulse of the u-concentration, the solution of the transport problem is (Bear, 1972)

$$u(x,y,t) = u_0 + \frac{\Delta u_0}{4\pi t\sqrt{D_L D_T}} exp\left(-\frac{\rho^2}{2}\right), \tag{63}$$

$$\rho(x,y,t) = \sqrt{\frac{(x - x_0 - vt)^2}{2D_L t} + \frac{(y - y_0)^2}{2D_T t}}, \tag{64}$$

where D_L and D_T are the longitudinal and transverse dispersion coefficient, and v is the groundwater velocity. From Eq. (63), the analytical solution of the mixing term in Eq. (58) is

$$f_{mix}(u) = \left(\frac{\Delta u_0}{4\pi t\sqrt{D_L D_T}}\right)^2 \frac{\rho^2}{2t} exp\left(-\rho^2\right). \tag{65}$$

This analytical solution was contrasted against the reconstructed solution of f_{mix} obtained from particle tracking simulations. Table 1 summarizes the parameters adopted for the numerical simulations. Transport simulations were performed in two steps: First, we used the RWPT Methodology described in Appendix B to simulate the migration of a passive solute representing the component u; Then, we evaluated f_{mix}, Eq. (65), by directly reconstructing the gradients of u using expression (45) along with Eqs. (46) and (47).

Transport simulations started with the injection of a large number of particles of equal mass, $m_p = \phi\Delta u_0/N_p$, at the point location (x_0,y_0), being N_p the total number of particles. Particle clouds were then measured at different times at which the reconstruction of u and f_{mix} was undertaken. The procedure used to reconstruct f_{mix} based on the proposed KDE method is described in Appendix C.

Figs. 1 and 2 compare the analytical solution of u and f_{mix} at time $t = 60$ with their corresponding optimal estimates obtained using the proposed KDE method. Simulations were performed with 2.5 million particles. The results are also contrasted against the traditional approach, which simply consists in counting particles within bins according to Eq. (12). In this approach, the gradients involved in f_{mix} were computed using a finite difference spatial discretization of u without any post-treatment. It is also compared to the coarse-graining technique, where the size of the bins was manually changed to obtain the most adequate smooth representation of u and f_{mix}. In overall, figures show that the proposed KDE method is capable to automate the reconstruction of concentrations and reactions rates from particle distributions, and further generate a more superior depiction of the reactions rates compared with the traditional solution. The latter is seen by noticing that, for the same particle distribution, the KDE method is capable to provide a more complete reproduction of the reaction rates than that of the traditional approach obtained after coarse-graining, which still requires adding more particles into the system. The main difference stems from the incomplete reproduction of the maximum values of the reaction rates given by the traditional approach.

The rate of convergence of the estimator of u and f_{mix} is shown in Fig. 3, which plots the coefficient of variation of the root mean square deviation of the estimator, *CV(RMSD)*, obtained with the traditional (without coarse-graining) and optimal approach as a function of the number of particles. Given some parameter p and its estimation $\hat{p}$, the *CV(RMSD)* was defined as

$$CV(RMSD) = \frac{RMSD}{\overline{p}}, \tag{66}$$

where

$$\overline{p} = \frac{1}{V}\int_V p(\mathbf{x})d\mathbf{x}, \tag{67}$$

$$RMSD = \left(\frac{1}{V}\int_V \left(p(\mathbf{x}) - \hat{p}(\mathbf{x})\right)^2 d\mathbf{x}\right)^{1/2}. \tag{68}$$

In applying the traditional approach, we fixed the size of the bin to $\Delta x = \Delta y = 1$. Results show that the rate of convergence associated with the KDE method is much smaller than that of the corresponding traditional method. Thus, the *CV(RMSD)* of concentrations displays a power law behavior with an exponent varying from −0.25 (KDE method) to −0.5 (traditional method). This is due to the fact that the use of an optimal support in the KDE method provides the "best unbiased" depiction of u and f_{mix} for any given number of particles, being the intersection between the traditional and optimal solution the number of particles needed in the traditional approach to obtain reliable predictions. Importantly, the traditional estimates of f_{mix} are orders of magnitude larger than its corresponding optimal solution, meaning that an unfeasible number of particles is required when using the traditional method.

Table 1
Flow and transport parameters adopted during the numerical simulations performed in a homogeneous medium.

Parameter	Value
Flow problem	
Number of cells in x direction, n_x	160
Number of cells in y direction, n_y	160
Cell size in x direction, $\Delta x[L]$	1.0
Cell size in y direction, $\Delta y[L]$	1.0
Reactive transport problem	
x coordinate of injection, x_0 [L]	20.0
y coordinate of injection, y_0 [L]	80.0
Injected mass, M_0 [M]	1.0
Darcy-Velocity, $q_x[L/T]$	0.3
Darcy-Velocity, $q_y[L/T]$	0.0
Porosity, $\phi[-]$	0.3
Longitudinal dispersion, $D_L[L^2/T]$	5.0
Transverse dispersion, $D_T[L^2/T]$	2.0
Equilibrium constant $[M^2/L^6]$	10^{-7}
Pulse injection of u, Δu_0 $[M/L^6]$	1.0
Initial concentration of u, u_0 $[M/L^6]$	0.0
Random walk features	
Grid Courant number [−]	0.1

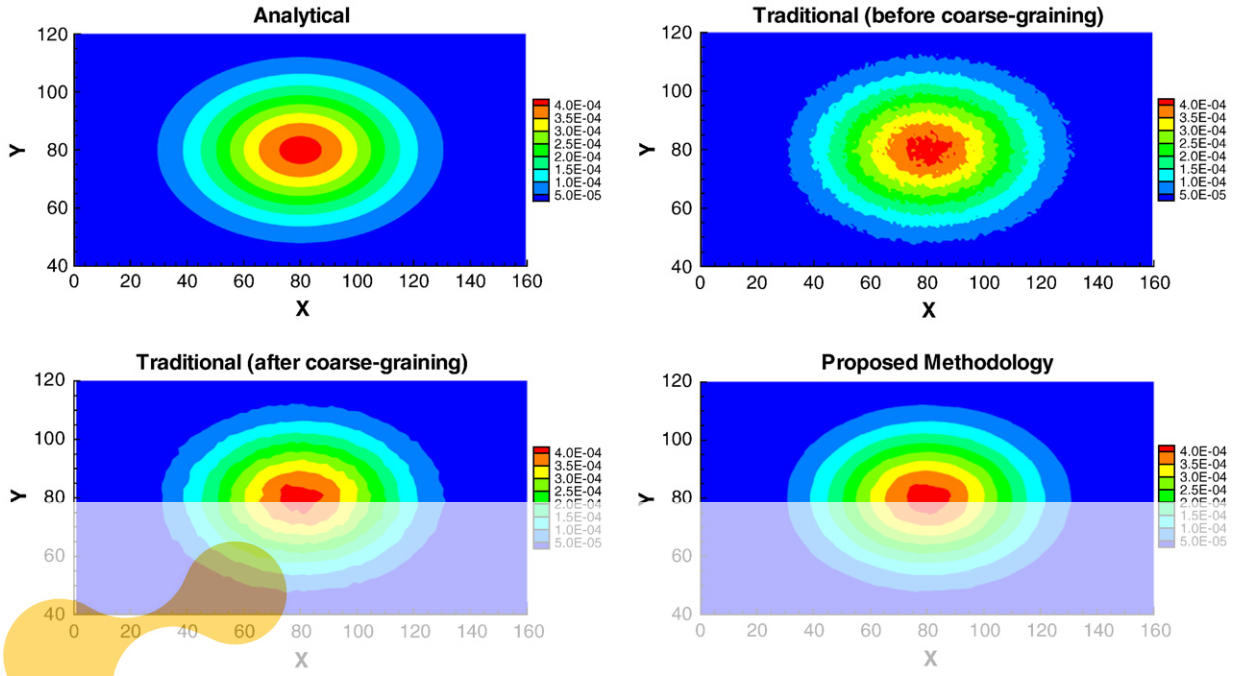

Fig. 1. Comparison of the analytical solution of u with the traditional approach (before and after coarse-graining) and the proposed KDE methodology. The bin sizes before and after coarse-graining were $\Delta x = \Delta y = 1.0$ unit and $\Delta x = \Delta y = 2.0$ units, respectively. Random walk simulations were performed with 2,500,000 particles.

Since the problem here is homogeneous and dispersive-dominated, the application of the traditional method satisfied that enough particles fall into bins of support size much smaller than the width of the mixing zone, where gradients are higher. Contrary to these desirable conditions, solute transport in the field is mostly advective-dominated and

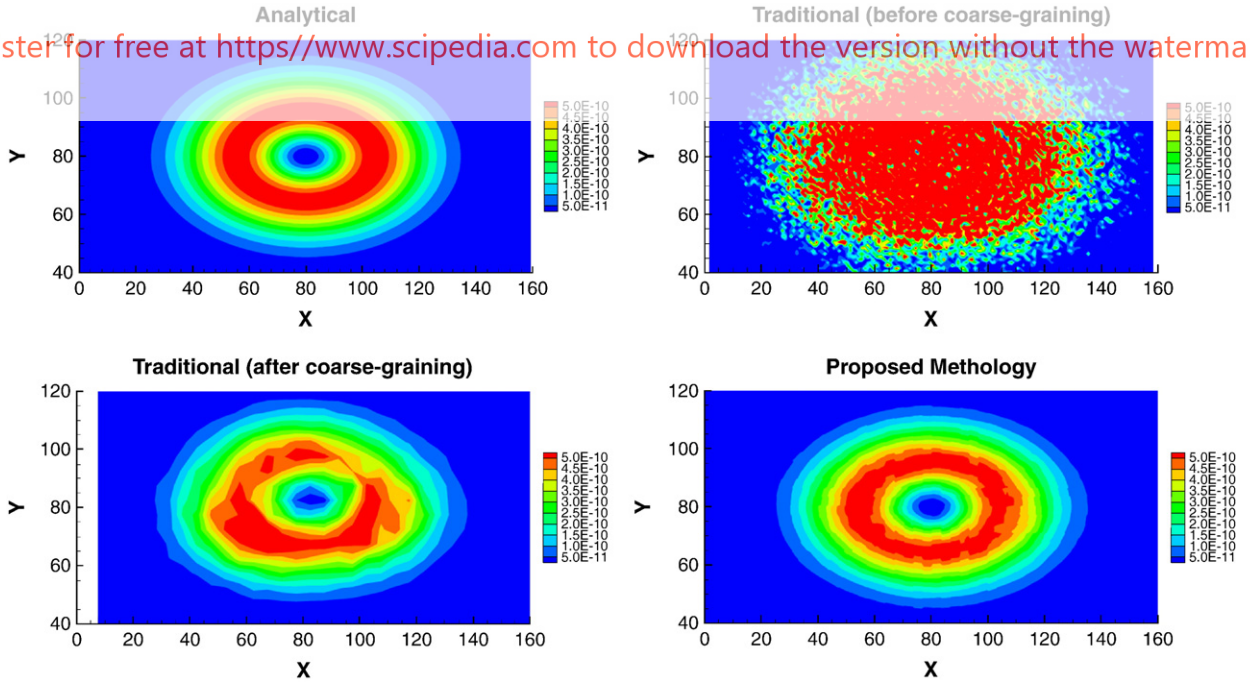

Fig. 2. Comparison of the analytical solution of f_{mix} with the traditional approach (before and after coarse-graining) and the proposed KDE methodology at a given time $t = 60$. The bin sizes before and after coarse-graining were $\Delta x = \Delta y = 1.0$ unit and $\Delta x = \Delta y = 4.5$ units, respectively. Random walk simulations were performed with 2,500,000 particles.

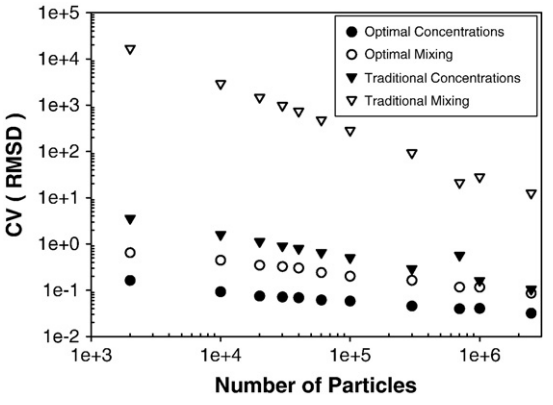

Fig. 3. Coefficient of variation of the root mean square deviation of the estimated concentration and mixing term obtained with the traditional and optimal approach as a function of the number of particles.

drastically influenced by heterogeneity. This produces highly distorted plumes with strong gradients and narrow mixing zones at the plume edges. Within this context, results indicate that the proposed KDE method to reconstruct reaction rates will allow the use of a more feasible number of particles to achieve the same quality of results.

4.3. Heterogeneous medium

A synthetic example of the effects of physical heterogeneity on reaction rates is presented in this section. For this purpose one realization of a sequential gaussian simulation of the natural log of the transmissivity field, $Y = lnT(\mathbf{x})$, of an aquifer was chosen. The following standardized isotropic spherical semivariogram was applied for the stochastic simulation of the transmissivity field

$$\frac{\gamma(h)}{\sigma_Y^2} = \begin{cases} 1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3 & \text{if } h \le a \\ 1 & \text{otherwise,} \end{cases} \quad (69)$$

where $a[L]$ is the range, h is the lag distance, and σ_Y^2 is the variance of lnT (see Table 2). The computational domain represents a squared aquifer with dimensions of $L_x = 160$ and $L_y = 160$, and a discretization of $\Delta x = \Delta y = 1.0$. The aquifer was assumed to be confined and with constant head boundaries at $x = 0$ and $x = 160$ and with the no-flow at the remaining model boundaries.

Two conceptual transport models were considered: Solute transport in Model A was purely influenced by advection and dispersion ($r_i \neq 0$ and $\Gamma_i = 0$), whereas in Model B a mass transfer equation was added to the advection–dispersion equation ($r_i \neq 0$ and $\Gamma_i \neq 0$). The parameters adopted for both transport models are summarized in Table 3. Again, transport simulations were conducted in two steps. First, we simulated the migration of the conservative species, u and u_{im}, using the RWPT methodology described in Appendix B; Then, we evaluated r_m and r_{im} through the expressions (57) and (58) by directly reconstructing the gradients of u and u_{im} using the proposed KDE method.

In both models, the aquifer was initially in geochemical equilibrium at all points. A water with a different chemical composition was then injected instantaneously in a rectangular area A_0 of 30 units width and 50 units height located orthogonal to the principal flow direction. A sketch of the set-up adopted for the transport simulations is provided in Fig. 4. Knowing again that $u = c_1 - c_2$ and $u_{im} = c_{1,im} - c_{2,im}$ are the conservative species for this problem, the initial condition can be written in terms of u and u_{im} as

Model A:

$$u\left(x,y,t=0^+\right) = \frac{M_0}{\phi_{tot} A_0} I\{x \in A_0\}, \quad (70)$$

Model B:

$$u\left(x,y,t=0^+\right) = \frac{M_0}{\phi_m A_0} I\{x \in A_0\}, \quad (71)$$

$$u_{im}\left(x,y,t=0^+\right) = 0. \quad (72)$$

Eq. (72) expresses that the injection of water takes place through the mobile zone (preferential flow paths). Parameters were chosen so that both plumes move at the same mean velocity and respond to the same impulse of u. The same mass was assigned to all particles, which were initially injected uniformly inside A_0 (see Table 3). Concentrations and reaction rates were then reconstructed from particle clouds measured at different times using the proposed KDE method.

The simulation results of u and r associated with model A are shown in Fig. 5 for a given time $t = 4064$. The two contributing terms involved in the reaction rate, i.e., mixing

Table 2
Flow problem parameters adopted during the numerical simulations performed in a heterogeneous medium.

Parameter	Value
Flow problem	
Number of cells in x direction, n_x	160
Number of cells in y direction, n_y	160
Cell size in x direction, $\Delta x[L]$	1.0
Cell size in y direction, $\Delta y[L]$	1.0
Mean hydraulic gradient in x direction, $J_x[-]$	0.001
Mean hydraulic gradient in y direction, $J_y[-]$	0.0
Heterogeneous field	
Geometric mean of T $[L^2/T]$	1.0
Variance of ln T $[-]$	3.6
Range of variogram $[L]$	16.0

Table 3
Transport parameters adopted during the numerical simulations performed in a heterogeneous medium.

Parameter	Model A	Model B
Area of injection, A_0 $[L^2]$	30×50	30×50
Injected mass, M_0 $[M]$	0.3	0.15
Mobile porosity, $\phi[-]$	0.3	0.15
Immobile porosity, $\phi[-]$	0.	0.15
Longitudinal dispersion, $D_L[L^2/T]$	0.3	0.3
Transverse dispersion, $D_T[L^2/T]$	0.1	0.1
Mass transfer coefficients, $\alpha_j[T^{-1}]$	0.	0.0002
Equilibrium constant $[M^2/L^6]$	10^{-7}	10^{-7}
Constant time step $[T]$	7.	7.
Number of particles $[-]$	2.5×10^6	5.0×10^6

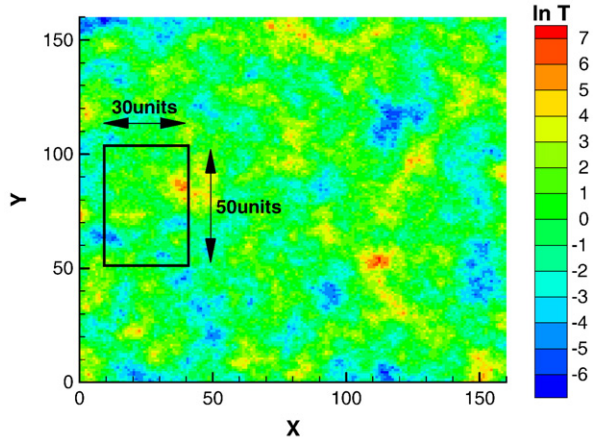

Fig. 4. Design of transport numerical simulations.

f_{mix} and chemical speciation f_{chem}, Eqs. (60) and (59), are also depicted. At this particular time, transverse dispersion has still not induced complete mixing, and the shape of the plume of the conservative component u is mostly distorted according to the velocity field. This produces sharp fronts and narrow mixing areas at the plume edges. Consistent with this picture, the proposed KDE method predicts chemical reactions to take place mostly at the plume boundaries. By looking at the mixing and chemistry terms, we see that in this particular case, the chemistry term mainly serves to amplify the mixing impulse without significantly changing the location of the hot spots of the reaction rates.

Most frequently in stochastic contaminant hydrology, the spatial fluctuations of the dispersion tensor due to the randomness of the velocity field are neglected and considered of minor importance. That is to say that the dispersion tensor in the conceptual transport model is assumed constant and fixed to some averaged value. This assumption may not be valid in reactive transport problems where the processes taking place at the local scale can still influence the global behavior (Fernàndez-Garcia et al., 2008). In order to assess the effects of assuming a constant dispersion during the calculation of the reaction rate, we estimated f_{mix} in two different ways:

$$\phi_m f_{mix}(\mathbf{x},t) = \alpha_L \|q(\mathbf{x},t)\| \left(\frac{\partial u}{\partial x}\right)^2 + \alpha_T \|q(\mathbf{x},t)\| \left(\frac{\partial u}{\partial y}\right)^2,$$

$$\phi_m f_{mix}(\mathbf{x},t) = \overline{D}_L \left(\frac{\partial u}{\partial x}\right)^2 + \overline{D}_T \left(\frac{\partial u}{\partial y}\right)^2, \tag{73}$$

where to make results comparable we have selected the constant diffusion parameters so that

$$\overline{D}_L = \alpha_L \overline{q} \qquad \overline{D}_T = \alpha_T \overline{q} \tag{74}$$

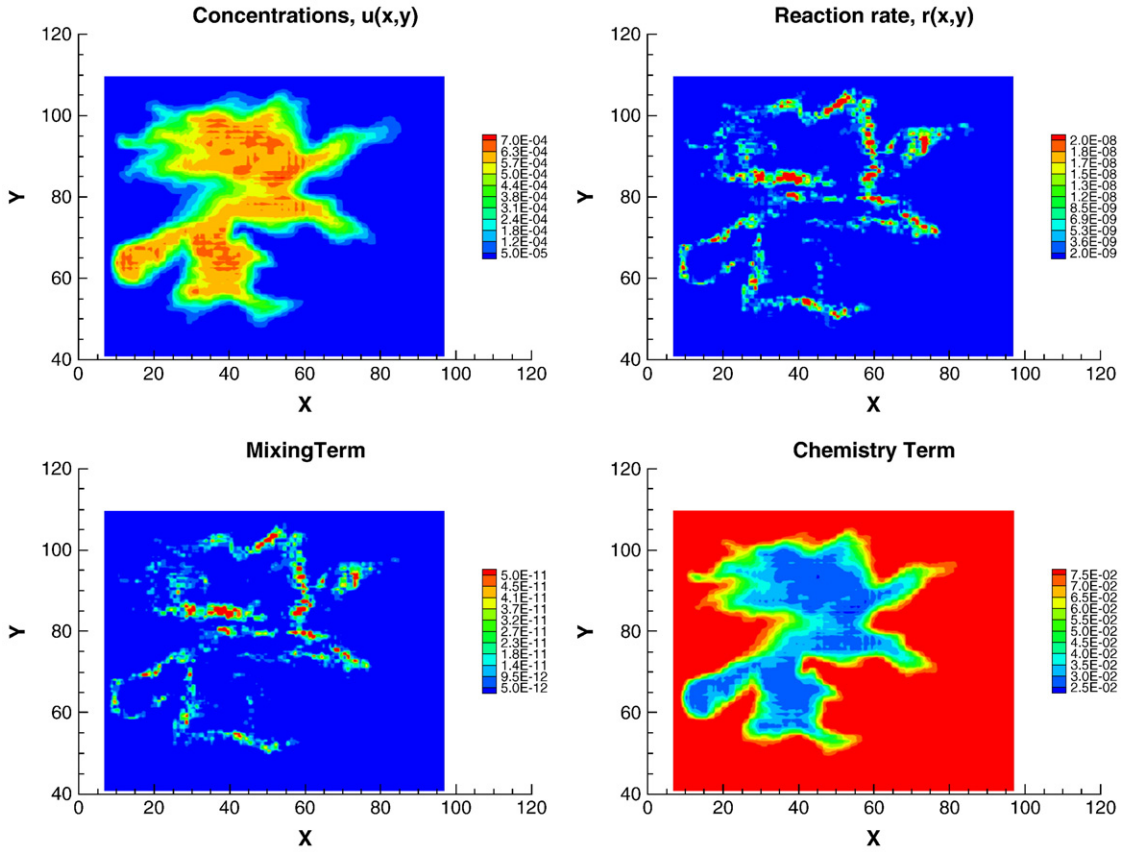

Fig. 5. Numerical simulations of concentrations and reactions rates in a heterogeneous aquifer at time $t = 4064$. Random walk simulations were performed with 2,500,000 particles.

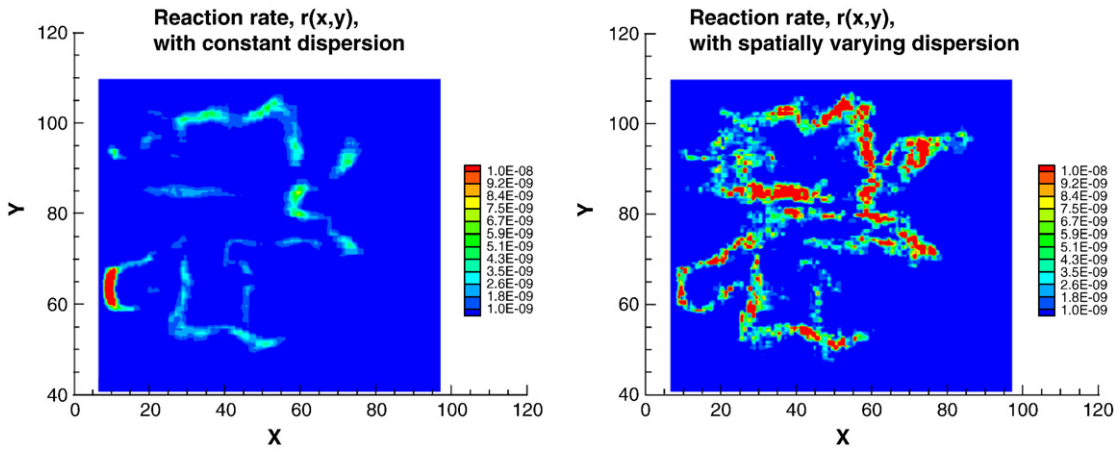

Fig. 6. Comparison of the reactions rates obtained in a heterogeneous aquifer using constant dispersion coefficients with those obtained using spatially varying dispersion coefficients at time $t = 4064$. The constant dispersion coefficients were computed using an average of the absolute value of the point velocities. Random walk simulations were performed with 2.5 million particles.

with

$$\overline{q} = \sqrt{\overline{q}_x^2 + \overline{q}_y^2}, \overline{q}_i = \frac{1}{V}\int_V |q_i(\mathbf{u})| d\mathbf{u}, i = 1, 2, \tag{75}$$

being V the area of the domain. The reaction rates simulated using a constant or a spatially varying dispersion tensor are shown in Fig. 6. Remarkably, a constant dispersion tensor significantly oversmoothed the reaction rates, concentrating their maximum values nearby the source where concentration gradients are higher. This is in contrast to the reaction rates obtained using a spatially varying dispersion tensor which showed larger reaction rates at the leading front of the plume. Recalling that mixing is caused by the joint effect of dispersion

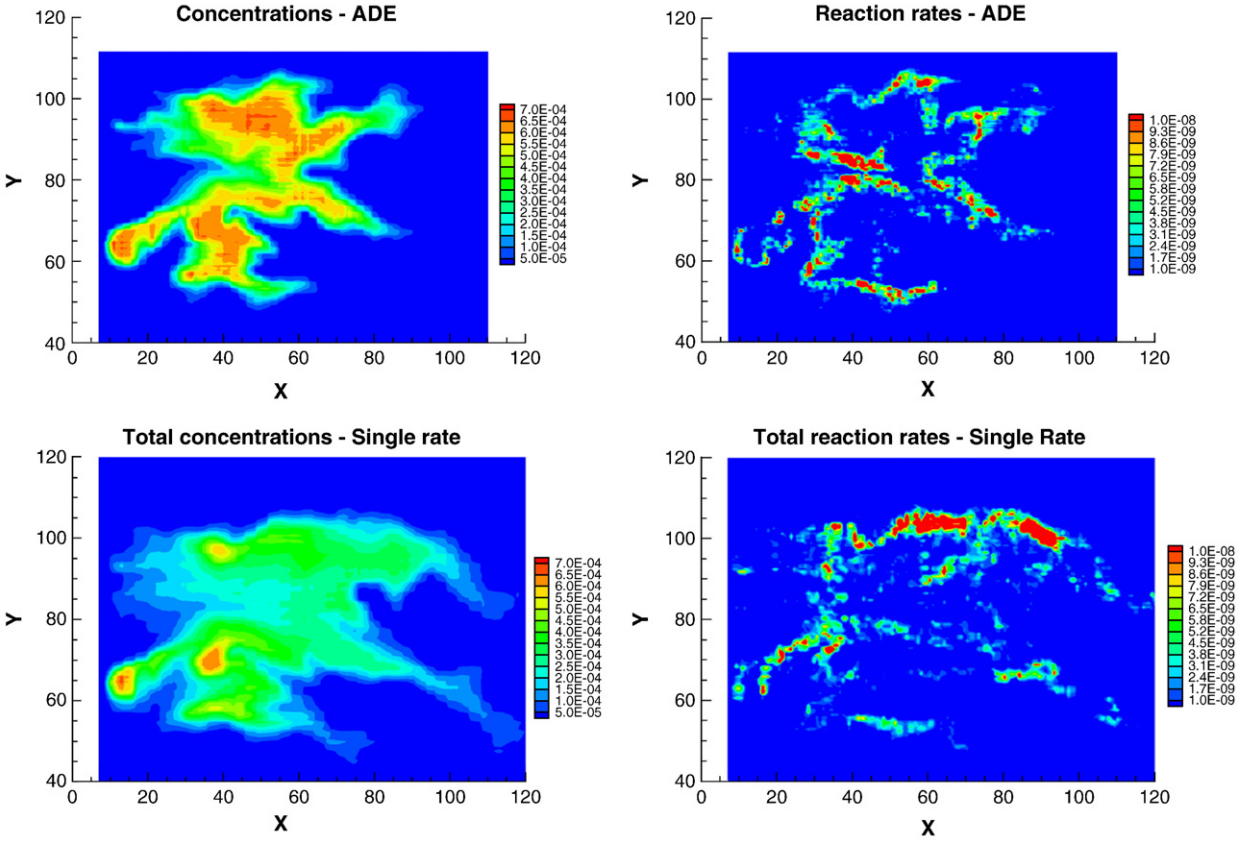

Fig. 7. Comparison of the reactions rates of the single-rate mass transfer model with the ADE model at time $t = 5410$. The total concentrations and reaction rates associated with the single-rate mass transfer model include both the mobile and the immobile contribution.

and concentration gradients, see Eq. (5), we attribute this behavior to the fact that the plume front is moving preferentially through areas of high velocities, where the effects of velocity-driven dispersion and large concentration gradients coexist. These effects are seen remarkably more important in the transverse direction to the mean flow direction.

The choice of a proper transport model may also drastically influence the model predictions of reaction rates. This is shown in Fig. 7, which compares the simulated reactions rates produced by models A and B at time $t=5410$. Interestingly, the addition of a mass transfer term into the advection–dispersion equation enhances the speed of dilution into the system. This is due to the joint effect of local mass transfer processes, intrinsically described by the transport model through $\Gamma_i(\mathbf{x},t)$ (Donado et al., 2009), and physical heterogeneity, i.e., the spatial variability of $q(\mathbf{x})$ in the governing transport equation.

5. Conclusions

The inherent fluctuations associated with the estimates of concentrations and their functionals (reaction rates) have typically prevented the use of particle tracking techniques to simulate complex reactive problems. In this paper, we have shown that this appreciation stems from a naive understanding of the reconstruction of functionals of concentrations in the particle tracking literature. Based on kernel density functions, we have presented an efficient method for the reconstruction of the reactions rates of chemical species from particle distributions. In doing this, the following main findings should be highlighted:

1. The reconstruction of reaction rates or other functionals of species concentrations in multi-dimensional reactive transport problems can easily be optimized and further automated via KDE methods. This provides a powerful tool for all branches of particle tracking techniques which require generating continuous fields of a system variable from discrete particle distributions. Thus, this method is well suited to directly reconstruct important system variables such as the dilution index E, the destruction of the concentration variance $f_{\sigma_c^2}$, the destruction/production of uncertainty f_{pdf}, the reaction rates of reactive species r, boundary mass fluxes and others.
2. Since the proposed KDE method avoids the propagation of the inherent fluctuations associated with the estimates of concentrations and their functionals, it renders particle tracking techniques the capability to be potentially coupled to any existing geochemical transport code. In this regard, we have shown that the coupling will be most efficient if the conceptual framework of De Simoni et al., 2005, which deconstructs the reactive problem into a conservative one plus speciation, is adopted to either calculate chemical species concentrations or their reaction rates.

The proposed KDE method was then used to analyze particle tracking simulations performed in a highly heterogeneous aquifer with the objective to evaluate the reactions rates of a bimolecular precipitation/dissolution chemical system. Two different transport conceptual models were considered: the traditional advection–dispersion model and the single-rate mass transfer model. The main results can be summarized as follows:

1. The proposed KDE method allows the reconstruction of sharp plume gradients and narrow mixing zones in heterogeneous aquifers, which are otherwise not well estimated by traditional approaches.
2. Simulation results show that the spatial variability of the dispersion tensor is crucial for making predictions of the reaction rates of chemical species taking place in heterogeneous aquifers. A constant dispersion tensor produces smoother representations of the reaction rate with a less amount of total precipitate. The reason is that the mixing term, f_{mix}, depends on the balance of two driven forces: velocity-driven dispersion and concentration gradients. At short times, the gradients close to the source are extremely large and control the reaction rate. As times evolve, the gradients slowly dissipate and compete with the velocity-driven dispersive processes. In this regime, the fact that the plume front moves preferentially through high velocities areas creates strong reaction rates at the leading edges of the plume. These effects were observed to be more important in the transverse direction to the mean flow direction.
3. The choice of the transport model can drastically effect mixing-driven chemical reactions. In this context, the joint effect of local mass transfer processes (intrinsically described in the transport model) and heterogeneity is shown to substantially enhanced mixing. At large times, when the plume is highly diluted (small concentration gradients), the reaction rate is mainly controlled by the transverse dispersion processes taking place at the leading front of the plume.

Acknowledgments

The authors acknowledge the financial support provided by ENRESA, by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Catalan Government, and by the Spanish CICYT through projects RARAVIS (CGL2009-11114), HEROS (CGL2007-66748), and SCARCE (CONSOLIDER).

Appendix A. Statistics of the estimators of concentrations and their derivatives

This Appendix describes the statistics of the estimates of univariate probability density functions and their derivatives. This is illustrated by actually looking at the estimates of $p(t)$, i.e., the normalized flux concentrations. Let us consider a random sample $T_1,...,T_{N_a}$ of N_a particle arrival times observed at a control location. Each random sample is described by the same probability density function $p(t)$, and assumed independent. This is a valid assumption in most random walk codes because the simulation of N_a particles starts from N_a independent uniformly (or normally) distributed numbers. We will also consider that all particles carry the same mass m_p. The kernel estimator of $\mathrm{d}^k p/\mathrm{d}t^k$ is

$$\hat{p}^{(k)}(t)\equiv\frac{1}{H^{1+k}M_a}\sum_p m_p K_k\left(\frac{t-T_p}{H}\right). \tag{76}$$

Optimal estimates of $p(t)$ can be determined by minimizing the Mean Integrated Square Error function (MISE), Eq. (22). MISE can be expressed as the sum of the integrated squared bias and the integral of variance through Eq. (23). We will start by evaluating the integrated squared bias. Taking the expected value of Eq. (76) we have that

$$E[\hat{p}(t)] = \frac{1}{H^{1+k}} E\left\{K_k\left(\frac{t-T_p}{H}\right)\right\}. \tag{77}$$

By the definition of the expected operator, this can be written as

$$E[\hat{p}(t)] = \frac{1}{H^{1+k}} \int K_k\left(\frac{t-\tau}{H}\right) p(\tau) d\tau, \tag{78}$$

Using then a change of variable $s=(t-\tau)/H$ yields

$$E[\hat{p}(t)] = \frac{1}{H^{k}} \int K_k(s) p(t-sH) ds, \tag{79}$$

which expresses that the expected value is given by the convolution of K_k with p. Using a Taylor expansion of p around t,

$$p(t+sH) = \sum_{n=0}^{\infty} \frac{1}{n!} p^{(n)}(t)(-1)^n s^n H^n, \tag{80}$$

and knowing that K_k satisfies Eq. (43) yields

$$E[\hat{p}(t)] = p^{(k)}(t) + \frac{H^{m-k}}{m!} p^{(m)}(t) \mu_m(K_k) + O\left(H^{m-k}\right), \tag{81}$$

so that the local squared bias of the estimator, $\left(E\left[\hat{p}^{(k)}\right]-p^{(k)}\right)^2$, is

$$Bias^2\left[\hat{p}^{(k)}(t)\right] = \frac{H^{2(m-k)}}{m!^2}\left(p^{(m)}(t)\right)^2 (\mu_m(K_k))^2 + O\left(H^{2(m-k)}\right). \tag{82}$$

On the other hand, the integral of the variance can be estimated as follows. Knowing that $M_a = N_p m_p$ and taking the variance of the estimator of $\hat{p}(t)$ we obtain

$$Var[\hat{p}(t)] = \frac{1}{H^{2(1+k)} N_p} Var\left\{K_k\left(\frac{t-T_p}{H}\right)\right\}, \tag{83}$$

where we have used that the random samples of particle arrival times are independent random variables. Knowing that $\mathrm{Var}[K_k] = E[K_k^2] - (E[K_k])^2$ and using the change of variable $s=(t-\tau)/H$ yields

$$Var\{K_k\} = \int K_k^2(s) p(t-sH) H ds - \left(\int K_k(s) p(t-sH) H ds\right)^2. \tag{84}$$

Using a Taylor expansion of $\hat{p}^{(k)}$ around t and knowing that K_k satisfies Eq. (43) yields

$$Var\{K_k\} = HR(K_k)p(t) - H^{2(1+k)}\left(p^{(k)}(t)\right)^2 + \ldots, \tag{85}$$

and therefore, substituting Eq. (85) into Eq. (83), leads to

$$Var[\hat{p}(t)] = \frac{1}{N_p H^{2k+1}}\left(R(K_k)p(t) - H^{2k+1}p^2(t)\right) + O\left(\left(N_p H^{2k+1}\right)^{-1}\right). \tag{86}$$

Substituting Eqs. (82) and (86) into Eq. (23), and taking the limit of *MISE* as $N_p H \to \infty$ and $H \to 0$, the following asymptotic value is obtained

$$AMISE(H) = \frac{H^{2(m-k)}}{m!^2} R\left(p^{(m)}(t)\right)(\mu_m(K_k))^2 + \frac{1}{N_p H^{2k+1}} R(K_k). \tag{87}$$

Appendix B. Algorithm to simulate transport with the random walk methodology

This Appendix describes the algorithm used to simulate the transport of the conservative species, u and u_{im}, with the random walk particle tracking methodology. The displacement of a particle is given by a drift term that related to the advective movement and a superposed Brownian motion responsable for dispersion,

$$\mathbf{X}_p(t+\Delta t) = \mathbf{X}_p(t) + \int_t^{t+\Delta t} \mathbf{A}\left(\mathbf{X}_p, \tau\right) d\tau + \mathbf{B}\left(\mathbf{X}_p, t\right) \Delta \mathbf{W}_t, \tag{88}$$

where Δt is the time step, $\mathbf{X}_p(t)$ is the position of a particle at time t, $\mathbf{A}$ is a drift vector, i.e., any change in $E[\mathbf{X}_p(t)]$ is due to the drift term, and $\mathbf{B}\triangle\mathbf{W}_\tau$ is the noise term. The tensor $\mathbf{B}$ is the displacement matrix that determines the strength of the particle random motion, and $\mathbf{W}_t$ is an n variable Wiener process determined by

$$\mathbf{B}\left(\mathbf{X}_p, t\right)\triangle\mathbf{W}_\tau = \mathbf{B}\left(\mathbf{X}_p, t\right)\xi(t)\sqrt{\Delta t}, \tag{89}$$

where $\xi(t)$ is a vector of independent (uncorrelated in space and time), normally distributed random variables with zero mean and unit variance.

Itô (1951) demonstrated that the particle density distribution $f(\mathbf{X}_p,t)$, defined as the probability of finding a particle within a given interval $[\mathbf{X}_p,\mathbf{X}_p+\mathrm{d}\mathbf{X}_p]$ at a given time t, obtained from Eq. (88) fulfills, in the limit of large particle numbers and an infinitesimally small step size, the Fokker–Planck equation. This partial differential equation is similar but not equal to the ADE. An analogy between them is established by the following relationship

$$\begin{aligned} \mathbf{A} &= \frac{\mathbf{q}}{\phi R} + \frac{1}{\phi R}\nabla\cdot(\phi\mathbf{D}) \\ 2\frac{\mathbf{D}}{R} &= \mathbf{B}\cdot\mathbf{B}^t. \end{aligned} \tag{90}$$

In the numerical simulations, the velocity field needed to simulate transport was obtained by previously solving the flow problem using a finite difference code, MODFLOW2000 (Harbaugh et al., 2000). This velocity field was then incorporated into

a random walk code, RW3D-MRMT (Fernàndez-Garcia et al., 2005; Salamon et al., 2006b) to solve the transport problem. In all cases, the same mass was associated to all particles, $m_p = M_0/N_p$, and the time integration of the drift term in Eq. (88) was calculated using the semi-analytical tracking method of Pollock (1988).

RW3D-MRMT simulates the single-rate mass transfer model by simply tracking the state of a particle (Salamon et al., 2006b). The state of a particle is an attribute that defines the domain at which the particle is present at a given time within the double porosity medium. The change from one state to another is easily determined using transition probabilities. The transition probability, $P_{ij}(\Delta t)$, that a particle presently in state i will be in state j at a time $t+\Delta t$ is given by

$$\mathbf{P}(\Delta t) = \begin{pmatrix} \dfrac{1+\beta e^{-(1+\beta)\alpha\Delta t}}{1+\beta} & \dfrac{1-e^{-(1+\beta)\alpha\Delta t}}{1+\beta} \\ \dfrac{\beta-\beta e^{-(1+\beta)\alpha\Delta t}}{1+\beta} & \dfrac{\beta+e^{-(1+\beta)\alpha\Delta t}}{1+\beta} \end{pmatrix}. \quad (91)$$

If a particle is in the mobile domain, then it is susceptible to advection and dispersion, otherwise the particle is not allowed to move. Having calculated the phase transition probabilities, numerical implementation into particle tracking is done easily. For each time step a uniform [0,1] random number Y is drawn for each particle and is compared to the corresponding probability. The state of a particle being in the mobile phase is adjusted accordingly.

Appendix C. Algorithm to reconstruct mixing and reaction rates

The algorithm used to directly estimate f_{mix} and r from particle clouds representing u during numerical simulations proceeds as follows: (1) Discretized the x and y axes in bins $\{B_{x_i}\}$ and $\{B_{y_j}\}$; (2) Estimate the marginal densities $p(x)$ and $p(y)$ and their derivatives using univariate kernel estimators with an optimal support; (3) Estimate the conditional densities $p\left(x|y\in B_{y_j}\right)$ and $p(y|x\in B_{x_i})$ and their derivatives for all i,j using univariate kernel estimators with an optimal support; (4) Reconstruct the partial derivatives of the bivariate probability density function based on Eqs. (46) and (47); (5) Calculate the gradients of u by means of Eq. (45); and (6) Estimate f_{mix} and r by using Eqs. (60) and (61), respectively.

The optimal bandwidth of the estimates of the probability density function and their derivatives were computed through Eqs. (28) and (44) by employing Gaussian Kernel functions. In doing this, the plug-in method of (Engel et al., 1994) was chosen to estimate the corresponding L^2-norm, $R(p^{(m)})$, involved in Eqs. (28) and (44).

References

Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE TTrans. Signal Process. 50 (2), 174–188.

Bear, J., 1972. Dynamics of Fluids in Porous Media. Elsevier, New York.

Bellin, A., Tonina, D., 2007. Probability density function of non-reactive solute concentration in heterogeneous porous formations. J. Contam. Hydrol. 94 (1–2), 109–125.

Benson, D.A., Wheatcrat, S.W., Meerschaert, M.M., 2000. Application of a fractional advection–dispersion equation. Water Resour. Res. 36 (6), 1403–1412.

Berkowitz, B., Cortis, A., Dentz, M., Scher, H., 2006. Modeling non-fickian transport in geological formations as a continuous time random walk. Rev. Geophys. 44. doi:10.1029/2005RG000178 RG2003.

Carrera, J., Sánchez-Vila, X., Benet, I., Medina, A., Galarza, G., Guimerà, J., 1998. On matrix diffusion: formulations, solution methods and qualitative effects. Hydrogeol. J. 6, 178–190.

Chang, Ch., Ansari, R., 2005. Kernel particle filter for visual tracking. IEEE Signal. Process. Lett. 12 (3), 242–245.

Dentz, M., Berkowitz, B., 2003. Transport behavior of a passive solute in continuous time random walks and multirate mass transfer, *Water Resour.* Res 39 (5), 1111. doi:10.1029/2001WR001163.

De Simoni, M., Carrera, J., Sanchez-Vila, X., Guadagnini, A., 2005. A procedure for the solution of multicomponent reactive transport problems. Water Resour. Res. 41, W11410. doi:10.1029/2005WR004056.

Donado, L. D., X. Sanchez-Vila, M. Dentz, J. Carrera, and D. Bolster (2009), Multicomponent reactive transport in multicontinuum media, *Water Resour. Res.*, doi:10.1029/2008WR006823, in press.

Duong, T., Hazelton, M.L., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. Nonparametric Stat. 15 (1), 17–30.

Duong, T., Cowling, A., Koch, I., Wand, M.P., 2008. Feature significance for multivariate kernel density estimation. Comput. Stat. Data Anal. 52, 4225–4242.

Engel, J., Herrmann, E., Gasser, T., 1994. An iterative bandwidth selector for kernel estimation of densities and their derivatives. Nonparametric Stat. 4, 21–34.

Fernàndez-Garcia, D., Illangasekare, T.H., Rajaram, Harihar, 2005. Differences in the scale dependence of dispersivity estimated from temporal and spatial moments in chemically and physically heterogeneous porous media. Adv. Water Resour. 28, 745–759.

Fernàndez-Garcia, D., Sanchez-Vila, X., Guadagnini, A., 2008. Reaction rates and effective parameters in stratified aquifers. Adv. Water Resour. 31, 1364–1376.

Fernàndez-Garcia, D., Llerar-Meza, G., Gómez-Hernández, J.J., 2009. Upscaling transport with mass transfer models: mean behavior and propagation of uncertainty. Water Resour. Res. 45, W10411. doi:10.1029/2009WR007764.

Fiorotto, V., Caroni, E., 2002. Solute concentration statistics in heterogeneous aquifers for finite Peclet values. Transp. Porous Media 48 (3), 331–351.

Gillespie, D., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81 (25), 2340–2361.

Haggerty, R., Gorelick, S.M., 1995. Multiple-rate mass transfer for modeling diffusion and surface reactions in media with pore-scale heterogeneity. Water Resour. Res. 31 (10), 2383–2400.

Haggerty, R., McKenna, S.A., Meigs, L.C., 2000. On the late-time behaviour of tracer test breakthrough curves. Water Resour. Res. 36 (12), 3467–3479.

Harbaugh, A.W., Banta, E.R., Hill, M.C., McDonald, M.G., 2000. MODFLOW-2000. The U.S. Geological Survey Modular Ground-Water Model—user guide to modularization concepts and the ground-water flow process, Open-file Report. 00-92.

Hardle, W., 1990. Smoothing Techniques with Implementation in S. Springer-Verlag, New York.

Herrera, P.A., Massabó, M., Beckie, R., 2008. A meshless method to simulate solute transport in heterogeneous porous media. Adv. Water Resour. 32 (3), 413–429.

Itô, K., 1951. On Stochastic Differential Equations.Vol. 4. Am. Math. Soc, New York, pp. 289–302.

Jones, M.C., Marron, J.S., Sheater, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. 91, 401–407.

Kapoor, V., Gelhar, L.W., 1994. Transport in three-dimensionally heterogeneous aquifers. 1. Dynamics of concentration fluctuations. Water Resour. Res. 30 (6), 1775–1788.

Kapoor, V., Kitanidis, P.K., 1998. Concentration fluctuations and dilution in aquifers. Water Resour. Res. 34 (5), 1181–1193.

Kitanidis, P.K., 1994. The concept of the dilution index, Water Resour. Res., 30 (7), 2011–2026.

Marron, J.S., Nolan, D., 1989. Canonical kernels for density estimation. Stat. Probab. Lett. 7, 195–199.

Neuman, S.P., Tartakovsky, D.M., 2008. Perspective on theories of non-Fickian transport in heterogeneous media. Adv. Water Res. 5, 670–680. doi:10.1016/j.advwatres.2008.08.005.

Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc. 85 (409), 66–72.

Pollock, D.W., 1988. Semianalytical computation of path lines for finite difference models. Ground Water 26 (6), 743–750.

Riva, M., Guadagnini, A., Fernàndez-Garcia, D., Sanchez-Vila, X., Ptak, T., 2008. Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the Lauswiesen site. J. Contam. Hydrol. 101, 1–13.

Sain, S.R., 2002. Multivariate locally adaptive density estimation. Comput. Stat. Data Anal. 39, 165–186.
Salamon, P., Fernàndez-Garcia, D., Gómez-Hernández, J.J., 2006a. A review and numerical assessment of the random walk particle tracking method. J. Contam. Hydrol. 87, 277–305.
Salamon, P., Fernàndez-Garcia, D., Gómez-Hernández, J.J., 2006b. Modeling mass transfer processes using random walk particle tracking. Water Resour. Res. VOL.42, W11417. doi:10.1029/2006WR004927 2006.
Sanchez-Vila, X., Guadagnini, A., Fernàndez-Garcia, D., 2009. Conditional probability density functions of concentrations for mixing-controlled reactive transport in heterogeneous aquifers. Math. Geosci. 41, 323–351.
Silverman, B., 1986. W. Chapman and Hall, Density Estimation for Statistics and Data Analysis, London.
Simonoff, J., 1995. A simple, automatic and adaptive bivariate density estimator based on conditional densities. Stat. Comput. 5, 245–252.
Stoessel, D., Sagerer, G., 2006. Kernel particle filter for visual quality inspection from monocular intensity. In: Franke, K., Muller, K.R., Nickolay, B., Schafer, R. (Eds.), Proceedings of Pattern Recognition: Lecture notes in computer science. Springer-Verlag, Berlin, pp. 597–606. 4174.
Tartakovsky, A., Meakin, P., 2005. A smoothed particle hydrodynamics model for miscible flow in three-dimensional fractures and two-dimensional Rayleigh–Taylor instability. J. Comput. Phys. 207, 610–624.
Wand, M.P., Jones, M.C., 1993. Comparison of smoothing parametrizations in bivariate kernel density estimation. J. Am. Stat. Assoc. 88, 520–528.
Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput Statist 9, 97–116.
Willmann, M., Carrera, J., Sanchez-Vila, X., 2008. Transport upscaling in heterogeneous aquifers: what physical parameters control memory functions? Water Resour. Res. 44, W12437. doi:10.1029/2007WR006531.
Willmann, M., J. Carrera, X. Sanchez-Vila, O. Silva (2009), Coupling of mass transfer and reactive transport for non-linear reactions in heterogeneous media,, *Water Resour. Res.*, In Press.